Machine Learning for Skin Permeability Prediction: Random Forest and XGBoost Regression

Kevin Ita, Joyce Prinze
College of Pharmacy Touro University, Mare Island-Vallejo California, CA 94592, USA

Correspondence to:
Kevin Ita, PhD
College of Pharmacy
Touro University,
Mare Island-Vallejo
California
CA 94592
USA
Telephone: 1-707-638-5994
kita@touro.edu

keywords: skin permeability, random forest, transdermal, pandas, descriptors

Abstract

Being able to precisely quantify percutaneous penetration for chemicals is helpful for assessing the potential toxicity risks and the development of transdermal drug delivery systems. Methods that can quickly and easily estimate skin permeability (Kp) are being actively sought by researchers. The linear free energy relationship (LFER) developed by Abraham is a practical method for predicting Kp. The permeability coefficients and Abraham solute descriptor values for 175 organic compounds have been documented. After bringing in the necessary Python modules, we loaded the permeability data into a Pandas dataframe. We used the solute descriptors and permeability coefficient as our X and y variables, respectively. One of the most effective ensemble algorithms in machine learning is the random forest, which works on the premise that one should build several independent random trees and then use the average of their predictions to form inferences. Another method is the XGBoost regression. This is an iterative gradient descent-based ensemble learning algorithm where the XGBoost's primary learner, the decision tree, is used for integration. Because of its outstanding accuracy, efficiency, and adaptability, XGBoost has quickly become one of the most favored methods for developing predictive models. In this research project, we utilized a publicly available dataset to make skin permeability predictions using the random forest and XBoost regression techniques.

Introduction

Exposure to chemicals in the environment, such as those found in pesticides, industrial compounds, and personal care items, can have negative effects on the skin[1]. The ability to precisely quantify percutaneous penetration for chemicals is helpful for assessing the potential toxicity risks[1]. The development of dermal and transdermal drug delivery systems as well as cosmetics also relies on the capacity to anticipate the skin permeability of chemical substances rapidly and accurately[2]. Transdermal formulation development can be aided by a skin permeation prediction model that accounts for the effects of penetration enhancers and other excipients found in drug delivery vehicles[3]. Noninvasive, painless, and continuous drug administration is possible via transdermal drug delivery systems, which have been shown to promote patient compliance and decrease adverse effects[3-5]. The use of animals for testing is costly, time-consuming, and may raise ethical questions[6]. When biological evidence is insufficient to ascertain flux, regulatory bodies such as the US Environmental Protection Agency turn to prediction algorithms to estimate percutaneous penetration[1]. Quantitative structure-activity relationship (QSAR) methods for calculating permeability make use of connections between the permeability coefficient and lipophilicity and diffusivity descriptors[6]. Models for quantitative

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4415496

structure-permeability relationship (QSPR) studies use experimental data for skin permeability and construct quantitative correlations through statistical methods, relating skin permeability to chemical structure as described by physico-chemical properties and other structural descriptors[7-9].

$$LogKp = 2.72 + 0.71\log Koct - 0.0061\ MW \qquad \text{Equation 1}$$

Equation 1 is the formula for the Potts and Guy Model (PGM), the most widely used skin permeability mathematical model, where MW is the molecular weight, logK is the partition coefficient, and Log Kp is the permeability coefficient[7, 10]. It has been reported that the Potts-Guy model may be overestimating dermal flux by a factor of 10,000 because most chemicals in everyday life are absorbed under fixed finite dose settings, but the data used to build these models comes from diverse laboratories and is generated under infinite dose conditions[1]. Furthermore, the compounds possess limited chemical heterogeneity[1]. In addition to quantitative structure-activity relationship (QSPR) studies, molecular dynamics (MD) simulations have also been used for skin permeability prediction[1]. When compared to QSPR approaches, MD simulations are much more time-consuming, but they are not restricted to a particular application space[3]. Yet, when compared to experimental observations, theoretical predictions have certain clear advantages, such as saving money and addressing ethical questions[9]. As an added bonus, the models may help shed light on the mechanics of absorption[8]. Sun et al., used K-nearest-neighbor regression, single-layer networks, a blend of experts, and Gaussian processes to predict the skin permeability coefficients of several compounds [11].

When evaluating cutaneous exposure to chemicals, the permeability coefficient (Kp) is a key parameter[6]. Investigators have published data on the Kp of several organic compounds[12, 13]. Thus, methods that can quickly and easily estimate skin permeability are being actively sought by researchers[6, 14]. The linear free energy relationship (LFER) developed by Abraham is a practical method for predicting Kp for ions and ionic species[15]. After initially being used to study the characteristics of neutral molecules[16], this method was expanded by Abraham and Acree to encompass ions and ionic species as well[17]. An equation with $R^2 = 0.832$ and SD = 0.46 log units was found by correlating the log Kp values of 119 solutes at 37°C with Abraham descriptors[18]. In addition, the standard deviation (SD) of log Kp predictions for three groups of 60 compounds was 0.48 log units[18]. Hydrogen bond basicity of the solute, which reduces the permeability coefficient, and the volume of the solute, which raises the permeability coefficient were found to be the primary determinants of log Kp[18]. The work of Zhang et al. shows a correlation between Kp and a number of chemical properties, such as the excess molar refraction E (in units of (cm^3/mol)/10), the combined dipolarity/polarizability S, the overall solute hydrogen bond acidity and basicity A and B, and the McGowan's characteristic molecular volume V (in units of (cm3/mol)/10)[15]. Cations and anions are further described by the symbols J^+ and J^- [19]. Predicting skin permeability is difficult since the parameter K(p), which is connected mathematically to the physicochemical properties of penetrants, is demonstrated to be fundamentally non-linear[20].

Machine learning and artificial intelligence (ML/AI), previously considered black box approaches, are becoming more interpretable, as a result of the recent advances in eXplainable artificial intelligence (XAI)[21]. These computational techniques have attracted the interest of many pharmaceutical researchers, in particular in the fields of drug administration and skin permeation[22, 23]. Autonomous decision-making for targeted goals is made possible by AI's use of interpretive systems and software that can learn from input data[24]. Pattern-recognition algorithms are the backbone of machine learning[24]. The predictive power and mechanistic understanding of phenomena that ML approaches may provide are promising[23]. It has been shown that new machine learning approaches are more accurate predictors than older ones[25]. One of the most effective ensemble algorithms in machine learning is the random forest, which works on the premise that one should build several independent random trees and then use the average of their predictions to form inferences[26].

A random forest model is made up of many different decision trees, and it may be used for both classification and regression[27]. The output of a random forest model is the mean value of the predictions made by each individual tree when used for regression tasks[27]. Many regression issues have been solved with this method[26]. Due to its emphasis on prediction rather than explanation, the RF prediction algorithm has seen explosive growth in its adoption as a scientific method[28]. In order to reduce variation, the random forest (RF) aggregates many different decision trees into a single model[28]. RF models, or random forests, are ensembles of prediction trees, where a large number of trees are combined to form a "forest" that may be used to classify

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4415496

data in diverse ways[29]. The method combines adaptive closest neighbors and bagging into a single nonparametric tree-based ensemble approach for efficient data-adaptive inference[30].

An algorithm called eXtreme Gradient Boosting (XGBoost) can transform unreliable prediction models into reliable ones[25, 31]. As a more refined version of the gradient boosting algorithm[32], this method is cutting edge. For supervised learning tasks, it is a machine-learning method that is highly efficient, powerful, and scalable[32]. XGBoost is an iterative gradient descent-based ensemble learning algorithm where the XGBoost's primary learner, decision tree, is used for integration[33]. The algorithm repeatedly divides features so that a tree can be constructed[33]. Using the best branch effect feature and threshold, each decision tree finishes the split construction[33]. A variable's role in producing a prediction is quantified by its gain value in XGBoost, with a higher gain value indicating greater significance[34]. Because of its outstanding accuracy, efficiency, and adaptability, XGBoost has quickly become one of the most favored methods for developing predictive models[32]. This study will utilize a publicly available dataset to make skin permeability predictions using the random forest and XBoost regression techniques.

Methods

Data Collection

Naseem et al. (Additional File 1) [6] documented the permeability coefficients and Abraham solute descriptor values for 175 organic compounds. Missing values and outliers were looked for in the data set. The feature (independent) variables in this case were the solute descriptors E, S, A, B, and V in the Abraham model[6, 15, 35] and the dependent (target) variable was the permeability coefficient logKp. Pandas-based methods were used via a JupyterLab Notebook. After bringing in the necessary Python modules, we loaded our data into a dataframe. We used the solute descriptors and permeability coefficient as our X and y variables, respectively.

Data splitting and model training

To better grasp the non-linear connection between the skin permeability coefficient and the Abraham descriptors, both random forest and XGBoost regression approaches were employed. The easiest approach to define the Random Forest is as a collection of various regression trees [35]. A regression tree is created by successively dividing the data into a series of rectangles [35]. Each variable split is determined by minimizing a criterion (such as the residual sum of squares) [35]. The feature space is partitioned into two node-stored regions[35]. These nodes are subdivided further until a stopping requirement (such as a minimum number of observations in terminal nodes) is fulfilled [35]. The response variable is then predicted using the mean of each group's values [35]. We created a separate dataset for training and testing. We built random forest regression and XGBoost models and used them to analyze the data. In our model, 100 decision trees were used. For XGBoost, we used the XGBRegressor class with hyperparameters passed as arguments to optimize our model.

Regression and Optimization

Using a regression tree to drive the RF regression algorithm is known as bagging[36]. Bagging is then used to generate tree 1 from the data [36] Using the aforementioned procedure, Tree-2, Tree-3, and Tree-n are formed. A prediction is made by averaging the trees[36].. A prediction is made through the average of trees[36]. The randomization in the random forest method allows for improved generalization performance, faster learning and testing methods, and a higher degree of accuracy[37]. Since more complicated models capture more details of existing patterns and increase the likelihood of overfitting, the random forest algorithm includes hyperparameters that regulate the complexity of the model[38]. The data set was divided into a training dataset (consisting of 80% of the total) and a test dataset (20%) for the purposes of model training and testing. There were 100 trees, and their highest depth was 20. To find the best configuration for our model in XGBoost, we utilized the XGBRegressor class with the following hyperparameters as arguments: n_ estimators=1000, learning rate=0.10, subsample=0.5, colsample bytree=1, and max depth=5. To evaluate the prediction result, the R2 (coefficient of determination) value was used[36]:

$$R^2 = 1 - \frac{u}{v} \qquad \text{Equation 2}$$

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4415496

Where R^2, u , and v represent the coefficient of determination, the residual (predicted value-actual value) sum of squares, and the total sum of squares, respectively. That is, u is the residual sum of squares (RSS) value, which is the average of the squares of the residuals, and v is the total sum.

Actual and Predicted Values

Using a random forest regression model[37], we were able to determine the significance of each feature as a node-impurity reduction weighted by the chance that each node reaches the node. In order to calculate the node probability, we divided the number of samples that made it to the node by the total number of samples taken. Using random forest regression, we were able to obtain feature importance ratings. We looked at the feature-importance ratings to determine which characteristics are useful for estimating skin permeability. Mean square error (MSE) of the target values was used to measure the accuracy of the predictions.

Data visualization

Data visualization was performed using Seaborn, a Python package built upon matplotlib. For further examination of the variables and their connections, a correlation matrix and heat map were created. To better understand how the model's performance shifts as hyperparameters are tweaked, we have created an accuracy heat map. In the heat map, the accuracy value rises as the color intensity rises. Also, we charted and visualized five decision trees. The pydot, kdeplot, and plot tree packages were used to generate the results' graphical representation.

Results and Discussion:

In the random forest technique, the algorithm randomly selects the training dataset T (b) (b=1,...,B) to train each Classification and Regression tree (CART)[39]. Out-of-bag samples refer to the data that was omitted during the random selection process[39]. The random forest approach randomly selects features or input variables from the input variables or features (MP) for creating each CART tree [39]. On the basis of randomly picked T (b) and P characteristics, the optimal splitting for each CART tree is determined [39]. The resulting ensemble trees are marked as follows[39]:

$$\{\Phi \boldsymbol{T}_b, m | b = 1..., B\} \qquad \text{Equation 3}$$

where $\boldsymbol{T}$ is the training dataset bagging repeatedly (B times), Φ is the mapping function, $\Phi \mathrm{T}$ represents the CART decision tree model learned from dataset $\boldsymbol{T}$**.** In the regression method, the final results can be obtained by majority voting[39]:

$$\hat{\mathbf{Y}} = \Phi T, P(\boldsymbol{X}) = \frac{1}{B} \sum_{b=1}^{B} \Phi \boldsymbol{T}_b, m(\boldsymbol{X}) \qquad \text{Equation 4}$$

P represents the number of input variables (features), while X represents the input variables. Equation 4 suggests that the averaged prediction in a random forest is the one that minimizes the average squared error of the individual predictions obtained from each CART tree[39]. This produces a more accurate prediction than the sum of all individual predictions[39].

$$\Delta I(Node) = I(d_{incoming}) - P_L I(d_L) - P_R I(d_R) \qquad \text{Equation 5}$$

denote Gini impurity, the impurity of the dataset value received from the upper node, the left lower node impurity, and the right lower node impurity, respectively. Here, PL and PR are the weight of the left and right lower nodes, respectively. The decision tree's application of the Gini impurity principle enables the prediction of feature significance[39].

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4415496

Our studies revealed that random forest regression for our prediction had an accuracy of 80.8%. The mean absolute error was 0.25, the mean square error was 0.25,

Our study focused on the effect of Abraham descriptors (excess molar refraction [E], the combined dipolarity/polarizability [S], the overall solute hydrogen bond acidity and basicity [A and B], and the McGowan's characteristic molecular volume V on skin permeability LogKp. We mapped the relative importance of several features (Figure 1). From the Feature importance shown, the overall solute hydrogen bond basicity (described by the B descriptor) was the most important parameter (0.6) in the plot of feature importance, whereas dipolarity/polarizability combination (S) and electronegativity (E) were the least important features (0.1 each) (excess molar refraction).

Decision tree estimators in an RF are separate models that are fitted on their own[40]. In order to minimize the error term, a tree learns by determining the ideal splitting points for each independent variable to partition the training data[41]. This is done in accordance with a predetermined loss function that penalizes errors. To model complex data, a non-linear method called a prediction tree is used to divide the covariates into optimal splits until it finds partitions that allow for the most homogenous sub-nodes[29]. RF is an ensemble supervised learning technique that presents the collective response of numerous individual decision trees constructed using random sampling[38]. The model's output is highly sensitive to the number of trees, the number of input variables, and the size of the nodes[37]. We demonstrated that skin permeability coefficients can be predicted using a random forest model.

Analysis of correlations between different variables was also undertaken. Correlations between the dependent variable Log Kp and the independent variables (E, S, A, B, and V) are shown in the heat map (Figure 2). Colored cells in a correlation heatmap reflect data on a normally monochromatic scale, displaying a 2D correlation matrix between two discrete dimensions. Each row in the table represents a value in the first dimension, and each column represents a value in the second dimension. The intensity of the cell's color indicates how many data points agree with the given dimension. It can be seen that E, S, A were positively correlated with LogKp and that B, and V were negatively correlated with the dependent variable. The RF regression model's performance was confirmed by the mean absolute error (MAE= 0.40) and the mean squared error (MSE= 0.255). The correlation coefficients between two sets of data can be displayed in a table called a correlation matrix. The most strongly correlated pairs of variables can be identified using correlation matrices. It is also possible to use them to find hidden connections between factors. It is possible to see how strongly two numerical variables are connected by using a correlation heatmap. The random forest regression model established a statistically significant association between the total descriptors and the skin permeability coefficient. Figures 3 displays the predicted and actual values via random forest regression. Figure 5 shows the decision trees that made up the random forest.

XGBoost sequentially ensembles decision trees to minimize the model error using a gradient descent optimization algorithm[21]. The fundamental concept of this algorithm is to minimize the following objective function, which consists of the loss function and regularization terms[32]:

$$\mathrm{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \qquad \text{Equation 6}$$

Here, l is the loss function that represents the error between observed data y_i, and predicted data $\hat{y}_1$, f_t is the model of the t -th tree, and t is the iteration index during the optimization process[32]. The detail of regulation term $\Omega(\mathrm{f})$ can be expressed as[32]:

$$\Omega(\mathrm{f}) = \gamma T + \frac{1}{2}\lambda w^2 \qquad \text{Equation 7}$$

where T indicates the total number of tree leaves, γ and λ are penalty coefficients, and w is a vector containing each leaf's score[32]. The parameter in equation can be found during the training process; however, several hyperparameters such as and must be defined before the training process can begin[32]. Figures 4 shows the predicted and actual values via XGBoost regression. The obtained results revealed that XGBoost regression

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4415496

had a Root Mean Square Error (RMSE) and the coefficient of determination (R^2) of 0.483410 and 0.848935 respectively.

Conclusion:

The development of dermal and transdermal drug delivery systems as well as cosmetics also relies on the capacity to anticipate the skin permeability of chemical substances rapidly and accurately. Recently, significant strides have been made regarding the use of machine learning tools for predictions. In this project, we used random forest and XGBoost regression to predict skin permeability values for a publicly available dataset. The dataset shows the effect of molecular descriptors on skin permeability. It is anticipated that this approach will facilitate the development of transdermal drug delivery systems as well as aid in the risk assessment of chemical exposure.

1. Burli, A., et al., *Organic compounds percutaneous penetration in vivo in man: Relationship to mathematical predictive model.* Regulatory Toxicology and Pharmacology, 2020. **112**: p. 104614.
2. Zhang, K., M.H. Abraham, and X. Liu, *An equation for the prediction of human skin permeability of neutral molecules, ions and ionic species.* Int J Pharm, 2017. **521**(1-2): p. 259-266.
3. Lundborg, M., et al., *Predicting drug permeability through skin using molecular dynamics simulation.* J Control Release, 2018.
4. Ita, K. and M. Ukaoma, *Progress in the transdermal delivery of antimigraine drugs.* Journal of Drug Delivery Science and Technology, 2022. **68**: p. 103064.
5. Ita, K. and S. Ashong, *Percutaneous Delivery of Antihypertensive Agents: Advances and Challenges.* AAPS PharmSciTech, 2020. **21**(2): p. 56.
6. Naseem, S., Y. Zushi, and D. Nabi, *Development and evaluation of two-parameter linear free energy models for the prediction of human skin permeability coefficient of neutral organic chemicals.* Journal of Cheminformatics, 2021. **13**(1): p. 25.
7. Potts, R.O. and R.H. Guy, *Predicting skin permeability.* Pharm Res, 1992. **9**(5): p. 663-9.
8. Moss, G.P. and M.T.D. Cronin, *Quantitative structure–permeability relationships for percutaneous absorption: re-analysis of steroid data.* International Journal of Pharmaceutics, 2002. **238**(1): p. 105-109.
9. Tsakovska, I., et al., *Quantitative structure-skin permeability relationships.* Toxicology, 2017. **387**: p. 27-42.
10. Potts, R.O. and R.H. Guy, *A predictive algorithm for skin permeability: the effects of molecular size and hydrogen bond activity.* Pharm Res, 1995. **12**(11): p. 1628-33.
11. Sun, Y., et al., *The application of stochastic machine learning methods in the prediction of skin penetration.* Applied Soft Computing, 2011. **11**(2): p. 2367-2375.
12. Flynn, G., *Physiochemical determinants of skin absorption.* Principles of route-to-route extrapolation for risk assessment, 1990: p. 93-127.
13. Cheruvu, H.S., et al., *An updated database of human maximum skin fluxes and epidermal permeability coefficients for drugs, xenobiotics, and other solutes applied as aqueous solutions.* Data in Brief, 2022. **42**: p. 108242.
14. Waters, L.J. and X.L. Quah, *Predicting skin permeability using HuskinDB.* Scientific Data, 2022. **9**(1): p. 584.
15. Zhang, K., M.H. Abraham, and X. Liu, *An equation for the prediction of human skin permeability of neutral molecules, ions and ionic species.* International Journal of Pharmaceutics, 2017. **521**(1): p. 259-266.
16. Abraham, M.H., *Scales of solute hydrogen-bonding: their construction and application to physicochemical and biochemical processes.* Chemical Society Reviews, 1993. **22**(2): p. 73-83.
17. Abraham, M.H. and W.E. Acree, Jr., *Equations for the transfer of neutral molecules and ionic species from water to organic phases.* J Org Chem, 2010. **75**(4): p. 1006-15.
18. Abraham, M.H. and F. Martins, *Human Skin Permeation and Partition: General Linear Free-Energy Relationship Analyses.* Journal of Pharmaceutical Sciences, 2004. **93**(6): p. 1508-1523.
19. Abraham, M.H. and Y.H. Zhao, *Determination of Solvation Descriptors for Ionic Species: Hydrogen Bond Acidity and Basicity.* The Journal of Organic Chemistry, 2004. **69**(14): p. 4677-4685.
20. Moss, G.P., et al., *The application and limitations of mathematical modelling in the prediction of permeability across mammalian skin and polydimethylsiloxane membranes.* J Pharm Pharmacol, 2011. **63**(11): p. 1411-27.
21. Li, Z., *Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost.* Computers, Environment and Urban Systems, 2022. **96**: p. 101845.
22. Bannigan, P., et al., *Machine learning models to accelerate the design of polymeric long-acting injectables.* Nature Communications, 2023. **14**(1): p. 35.
23. Ashrafi, P., et al., *The application of machine learning to the modelling of percutaneous absorption: an overview and guide.* SAR QSAR Environ Res, 2015. **26**(3): p. 181-204.

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4415496

24. Paul, D., et al., *Artificial intelligence in drug discovery and development.* Drug Discov Today, 2021. **26**(1): p. 80-93.
25. Hou, N., et al., *Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost.* J Transl Med, 2020. **18**(1): p. 462.
26. Gao, W., F. Xu, and Z.-H. Zhou, *Towards convergence rate analysis of random forests for classification.* Artificial Intelligence, 2022. **313**: p. 103788.
27. Xia, Z. and K. Stewart, *A counterfactual analysis of opioid-involved deaths during the COVID-19 pandemic using a spatiotemporal random forest modeling approach.* Health & Place, 2023. **80**: p. 102986.
28. Couronné, R., P. Probst, and A.-L. Boulesteix, *Random forest versus logistic regression: a large-scale benchmark experiment.* BMC Bioinformatics, 2018. **19**(1): p. 270.
29. Austin, A.M., et al., *Using a cohort study of diabetes and peripheral artery disease to compare logistic regression and machine learning via random forest modeling.* BMC Med Res Methodol, 2022. **22**(1): p. 300.
30. Chen, X. and H. Ishwaran, *Random forests for genomic data analysis.* Genomics, 2012. **99**(6): p. 323-329.
31. Chen, T. and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*. 2016. 785-794.
32. Tarwidi, D., et al., *An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach.* MethodsX, 2023: p. 102119.
33. Dong, C., et al., *Non-contact screening system based for COVID-19 on XGBoost and logistic regression.* Computers in Biology and Medicine, 2022. **141**: p. 105003.
34. Khera, R., et al., *Use of Machine Learning Models to Predict Death After Acute Myocardial Infarction.* JAMA Cardiol, 2021. **6**(6): p. 633-641.
35. Liu, X., W.E. Acree, and M.H. Abraham, *Descriptors for some compounds with pharmacological activity; calculation of properties.* International Journal of Pharmaceutics, 2022. **617**: p. 121597.
36. Kwak, S., et al., *Machine learning prediction of the mechanical properties of γ-TiAl alloys produced using random forest regression model.* Journal of Materials Research and Technology, 2022. **18**: p. 520-530.
37. Nam, D.Y. and J.K. Rhee, *Assessment of MicroRNAs Associated with Tumor Purity by Random Forest Regression.* Biology (Basel), 2022. **11**(5).
38. Trindade, P.H.E., et al., *Improving Ovine Behavioral Pain Diagnosis by Implementing Statistical Weightings Based on Logistic Regression and Random Forest Algorithms.* Animals (Basel), 2022. **12**(21).
39. Xue, L., et al., *A data-driven shale gas production forecasting method based on the multi-objective random forest regression.* Journal of Petroleum Science and Engineering, 2021. **196**: p. 107801.
40. Dutschmann, T.M. and K. Baumann, *Evaluating High-Variance Leaves as Uncertainty Measure for Random Forest Regression.* Molecules, 2021. **26**(21).
41. Cattani, G., *Combining data envelopment analysis and Random Forest for selecting optimal locations of solar PV plants.* Energy and AI, 2023. **11**: p. 100222.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4415496